# Chapter 1
# Introduction and Motivating Examples

## 1.1 Introduction

This book examines how a response is related to covariates using mathematical models whose unknown parameters we wish to estimate using available information—this endeavor is known as *regression analysis*. In this first chapter, we will begin in Sect. 1.2 by making some general comments about model formulation. In Sect. 1.3, a number of examples will be described in order to motivate the material to follow in the remainder of this book. In Sect. 1.4, we examine, in simple idealized scenarios, how "randomness" is induced by not controlling for covariates in a model. Section 1.5 briefly contrasts the Bayesian and frequentist approaches to inference, and Sect. 1.7 gives references that expand on the material of this chapter. Finally, Sect. 1.6 summarizes the overall message of this book which is that in many instances, carefully thought out Bayesian and frequentist analyses will provide similar conclusions; however, situations in which one or the other approach may be preferred are also described.

## 1.2 Model Formulation

In a regression analysis, the following steps may be followed:

1. Formulate a model based on the nature of the data, the subject matter context, and the aims of the data analysis.
2. Examine the mathematical properties of the initial model with respect to candidate inference procedures. This examination will focus on whether specific methods are suited to both the particular context under consideration and the specific questions of interest in the analysis.
3. Consider the computational aspects of the model.

J. Wakefield, *Bayesian and Frequentist Regression Methods*, Springer Series
in Statistics, DOI 10.1007/978-1-4419-0925-1_1,

The examination in steps 2 and 3 may suggest that we need to change the model.[1] Historically, the range of model forms that were available for regression modeling was severely limited by computational and, to a lesser extent, mathematical considerations. For example, though *generalized linear models* contain a flexible range of alternatives to the linear model, a primary motivation for their formulation was ease of fitting and mathematical tractability. Hence, step 3 in particular took precedent over step 1.

Specific aspects of the initial model formulation will now be discussed in more detail. When carrying out a regression analysis, careful consideration of the following issues is vital and in many instances will outweigh in importance the particular model chosen or estimation method used. The interpretation of parameters also depends vitally on the following issues.

### *Observational Versus Experimental Data*

An important first step in data analysis is to determine whether the data are experimental or observational in nature. In an experimental study, the experimenter has control over at least some aspects of the study. For example, units (e.g., patients) may be randomly assigned to covariate groups of interest (e.g., treatment groups). If this randomization is successfully implemented, any differences in response will (in expectation) be due to group assignment only, allowing a causal interpretation of the estimated parameters. The beauty of randomization is that the groups are balanced with respect to all covariates, crucially including those that are *unobserved*.

In an observational study, we never know whether observed differences between the responses of groups of interest are due, at least partially, to other "confounding" variables related to group membership. If the confounders are measured, then there is some hope for controlling for the variability in response that is not due to group membership, but if the confounders are unobserved variables, then such control is not possible. In the epidemiology and biostatistics literature, this type of discrepancy between the estimate and the "true" quantity of interest is often described as bias due to confounding. In later chapters, this issue will be examined in detail, since it is a primary motivation for regression modeling. In observational studies, estimated coefficients are traditionally described as *associations*, and causality is only alluded to more informally via consideration of the combined evidence of different studies and scientific plausibility. We expand upon this discussion in Sect. 1.4.

Predictive models are more straightforward to build than causal models. To quote Freedman (1997), "For description and prediction, the numerical values of the individual coefficients fade into the background; it is the whole linear combination on the right-hand side of the equation that matters. For causal inference, it is the individual coefficients that do the trick."

[1] To make clear, we are not suggesting refining the model based on inadequacies of fit; this is a dangerous enterprise, as we discuss in Chap. 4.

## *Study Population*

Another important step is to determine the population from which the data were collected so that the individuals to whom inferential conclusions apply may be determined. Extrapolation of inference beyond the population providing the data is a risky enterprise.

Throughout this book, we will take a superpopulation view in which probability models are assumed to describe variability with respect to a hypothetical, infinite population. The study population that exists in practice consists of $N$ units, of which $n$ are sampled. To summarize:

$$\text{Superpopulation } (\infty) \quad \rightarrow \quad \text{Study Population } (N) \quad \rightarrow \quad \text{Sample } (n)$$

Inference for the parameters of a superpopulation may be contrasted with a survey sampling perspective in which the focus is upon characteristics of the responses of the $N$ units; in the latter case, a full census ($n = N$) will obviate the need for statistical analysis.

## *The Sampling Scheme*

The data collection procedure has implications for the analysis, in terms of the models that are appropriate, the questions that may be asked, and the inferential approach that may be adopted. In the most straightforward case, the data arise through random sampling from a well-defined population. In other situations, the random samples may be drawn from within covariate-defined groups, which may improve efficiency of estimation by concentrating the sampling in informative groups but may limit the range of questions that can be answered by the data due to the restrictions on the sampling scheme. In more complex situations, the data may result from outcome-dependent sampling. For example, a case-control study is an outcome-dependent sampling scheme in which the binary response of interest is fixed by design, and the random variables are the covariates sampled within each of the outcome categories (cases and controls). For such data, care is required because the majority of conventional approaches will not produce valid inference, and analysis is carried out most easily using logistic regression models. Similar issues are encountered in the analysis of matched case-control studies, in which cases and controls are matched upon additional (confounder) variables. Bias in parameters of interest will occur if such data are analyzed using methods for unmatched studies, again because the sampling scheme has not been acknowledged. In the case of individually matched cases and controls (in which, for example, for each case a control is picked with the same gender, age, and race), conventional likelihood-based methods are flawed because the number of parameters (including one parameter for each case-control pair) increases with the sample size (providing an example of the importance of paying attention to the regularity conditions required for valid inference)—*conditional* likelihood provides a valid inferential approach in this case. The analysis of data from case-control studies is described in Chap. 7.

## *Missing Data*

Measurements may be missing on the responses which can lead to bias in estimation, depending on the reasons for the absence. It is clear that bias will arise when the probability of missingness depends on the size of the response that would have been observed. An extreme example is when the result of a chemical assay is reported as "below the lower limit of detection"; such a variable may be reported as the (known) lower limit, or as a zero, and analyzing the data using these values can lead to substantial bias. Removing these observations will also lead to bias. In the analysis of individual-level data over time (to give so-called longitudinal data) another common mechanism for missing observations is when individuals drop out of the study.

## *Aim of the Analysis*

The primary aim of the analysis should always be kept in mind; in particular, is the purpose descriptive, exploratory (e.g., for hypothesis generation), confirmatory (with respect to an a priori hypothesis), or predictive? Regression models can be used for each of these endeavors, but the manner of their use will vary. Large data sets can often be succinctly described using parsimonious[2] regression models. Exploratory studies are often informal in nature, and many different models may be fitted in order to gain insights into the structure of the data. In general, however, great care must be taken with data dredging since spurious associations may be discovered due to chance alone.

The level of sophistication of the analysis, and the assumptions required, will vary as the aims and abundance of data differ. For example, if one has a million observations independently sampled from a population, and one requires inference for the mean of the population, then inference may be based on the sample mean and sample standard deviation alone, without recourse to more sophisticated models and approaches—we would expect such inference to be reliable, being based on few assumptions. Similarly, inference is straightforward if we are interested in the average response at an observed covariate value for which abundant data were recorded.

[2] The Oxford English Dictionary describes *parsimony* as "... that no more causes or forces should be assumed than are necessary to account for the facts," which serves our purposes, though care is required in the use of the words "causes," "forces," and "facts."

However, if such data are not available (e.g., when the number of covariates becomes large or the sample size is small), or if interpolation is required, regression models are beneficial, as they allow the totality of the data to estimate global parameters and smooth across unstructured variability. To answer many statistical questions, very simple approaches will often suffice; the *art* of statistical analysis is deciding upon when a more sophisticated approach is necessary/warranted, since dependence on assumptions usually increases with increasing sophistication.

## 1.3 Motivating Examples

We now introduce a number of examples to illustrate different data collection procedures, types of data, and study aims. We highlight the distinguishing features of the data in each example and provide a signpost to the chapter in which appropriate methods of analysis may be found.

In general, data $\{Y_i, \boldsymbol{x}_i, i = 1, \ldots, n\}$ will be available on $n$ units, with $Y_i$ representing the univariate response variable and $\boldsymbol{x}_i = [1, x_{i1}, \ldots, x_{ik}]$ the row vector of explanatory variables on unit $i$. Variables written as uppercase letters will represent random variables, and those in lowercase fixed quantities, with boldface representing vectors and matrices.

### *1.3.1 Prostate Cancer*

We describe a dataset analyzed by Tibshirani (1996) and originally presented by Stamey et al. (1989). The data were collected on $n = 97$ men before radical prostatectomy, which is a major surgical operation that removes the entire prostate gland along with some surrounding tissue. We take as response, $Y$, the log of prostate specific antigen (PSA); PSA is a concentration and is measured in ng/ml. In Stamey et al. (1989), PSA was proposed as a preoperative marker to predict the clinical stage of cancer. As well as modeling the stage of cancer as a function of PSA, the authors also examined PSA as a function of age and seven other histological and morphometric covariates. We take as our aim the building of a predictive model for PSA, using the eight covariates:

- log(can vol): The log of cancer volume, measured in milliliters (cc). The area of cancer was measured from digitized images and multiplied by a thickness to produce a volume.
- log(weight): The log of the prostate weight, measured in grams.
- Age: The age of the patient, in years.
- log(BPH): The log of the amount of benign prostatic hyperplasia (BPH), a noncancerous enlargement of the prostate gland, as an area in a digitized image and reported in cm$^2$.

- SVI: The seminal vesicle invasion, a 0/1 indicator of whether prostate cancer cells have invaded the seminal vesicle.
- log(cap pen): The log of the capsular penetration, which represents the level of extension of cancer into the capsule (the fibrous tissue which acts as an outer lining of the prostate gland). Measured as the linear extent of penetration, in cm.
- Gleason: The Gleason score, a measure of the degree of aggressiveness of the tumor. The Gleason grading system assigns a grade (1–5) to each of the two largest areas of cancer in the tissue samples with 1 being the least aggressive and 5 the most aggressive; the two grades are then added together to produce the Gleason score.
- PGS45: The percentage of Gleason scores that are 4 or 5.

The BPH and capsular penetration variables originally contained zeros, and a small number was substituted before the log transform was taken. It is not clear from the original paper why the log transform was taken though PSA varies over a wide range, and so linearity of the mean model may be aided by the log transform. It is also not clear why the variable PGS45 was constructed. If initial analyses were carried out to find variables that were associated with PSA, then significance levels of hypothesis tests will not be accurate (since they are not based on an a priori hypotheses but rather are the result of data dredging).

Carrying out exploratory data analysis (EDA) is a vital step in any data analysis. Such an enterprise includes the graphical and tabular examination of variables, the checking of the data for errors (for example, to see if variables are within their admissible ranges), and the identification of outlying (unusual) observations or influential observations that when perturbed lead to large changes in inference. This book is primarily concerned with methods, and the level of EDA that is performed will be less than would be desirable in a serious data analysis.

Figure 1.1 displays the response plotted against each of the covariates and indicates a number of associations. The association between $Y$ and log(can vol) appears particularly strong. In observational settings such as this, there are often strong dependencies between the covariates. We may investigate these dependencies using scatterplots (or tables, if both variables are discrete). Figure 1.2 gives an indication of the dependencies between those variables that exhibit the strongest associations; log(can vol) is strongly associated with a number of other covariates. Consequently, we might expect that adding log(can vol) to a model for log(PSA) that contains other covariates will change the estimated associations between log(PSA) and the other variables.

We define $Y_i$ as the log of PSA and $\boldsymbol{x}_i = [1, x_{i1}, \ldots, x_{i8}]$ as the $1 \times 9$ row vector associated with patient $i$, $i = 1, \ldots, n = 97$. We may write a general mean model as $\mathrm{E}[Y_i \mid \boldsymbol{x}_i] = f(\boldsymbol{x}_i, \boldsymbol{\beta})$ where $f(\cdot, \cdot)$ represents the functional form and $\boldsymbol{\beta}$ unknown regression parameters. The most straightforward form is the multiple linear regression

$$f(\boldsymbol{x}_i, \boldsymbol{\beta}) = \beta_0 + \sum_{j \in C} x_{ij} \beta_j, \tag{1.1}$$

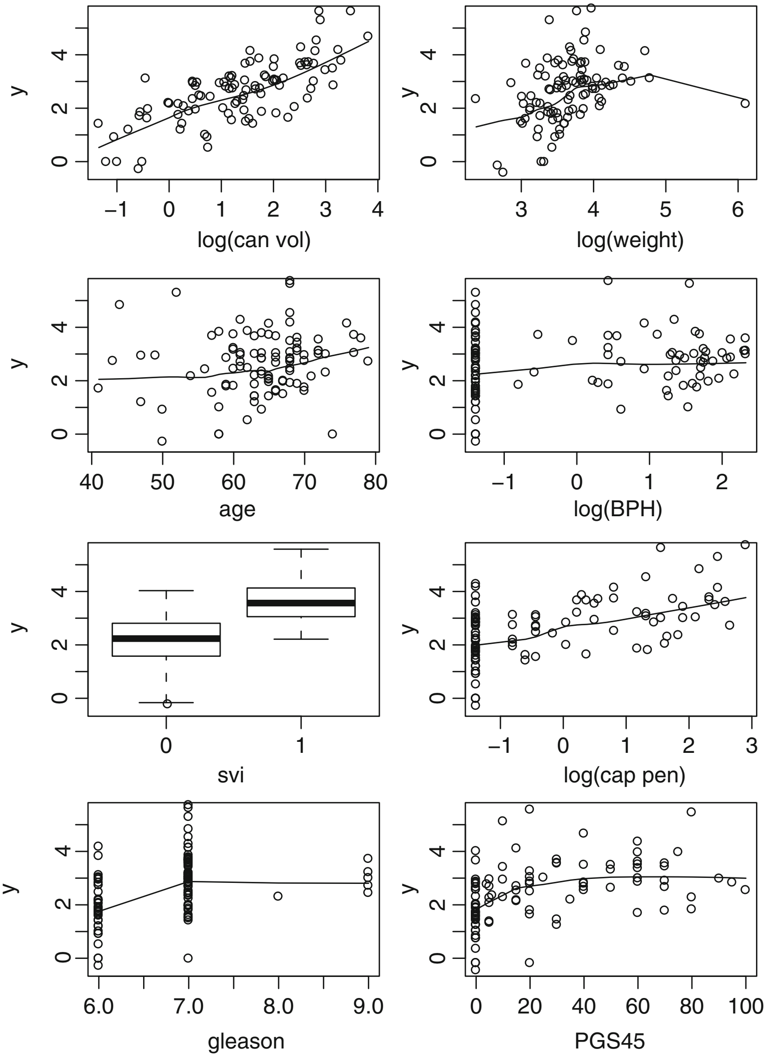

**Fig. 1.1** The response $y = \log(\mathrm{PSA})$ plotted versus each of the eight explanatory variables, $x$, in the prostate cancer study, with local smoothers superimposed for continuous covariates

where $C$ corresponds to the subset of elements of $\{1, 2, \ldots, 8\}$ whose associated covariates we wish to include in the model and $\boldsymbol{\beta} = [\beta_0, \{\beta_j, j \in C\}]^{\mathrm{T}}$. The interpretation of each of the coefficients $\beta_j$ depends crucially on knowing the scaling and units of measurement of the associated variables $x_j$.

Most of the $x$ variables in this study are measured with error (as is clear from their derivation, e.g., log(BPH) is derived from a digitized image), and if we are interested in estimating causal effects, then this aspect needs to be acknowledged

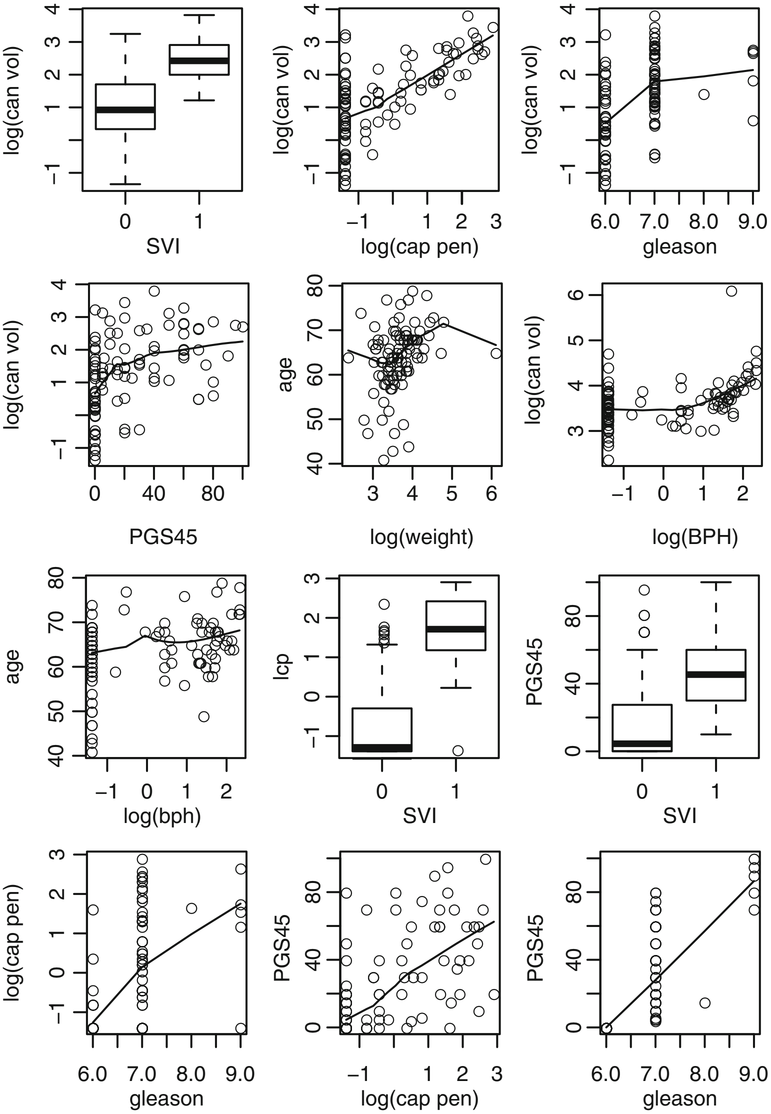

**Fig. 1.2** Associations between selected explanatory variables in the prostate cancer study, with local smoothers superimposed for continuous covariates

in the models that are fitted, since inference is affected in this situation, which is known as *errors-in-variables*.

*Distinguishing Features.* Inference for multiple linear regression models is described in Chap. 5, including a discussion of parameter interpretation. Chapter 4 discusses the difficult but important topics of model formulation and selection.

**Table 1.1** Outcome after head injury as a function of four covariates: pupils, hematoma present, coma score, and age

| | Pupils | | Good | | | | Poor | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Hematoma present | | No | | Yes | | No | | Yes | |
| | Coma score | Low | High | Low | High | Low | High | Low | High | |
| | 1–25 | Dead | 9 | 5 | 5 | 7 | 58 | 11 | 32 | 12 |
| | | Alive | 47 | 77 | 11 | 24 | 29 | 24 | 13 | 16 |
| Age | 26–54 | Dead | 19 | 6 | 21 | 14 | 45 | 7 | 61 | 15 |
| (years) | | Alive | 15 | 44 | 18 | 38 | 11 | 16 | 11 | 21 |
| | $\geq$55 | Dead | 7 | 12 | 19 | 25 | 20 | 7 | 42 | 17 |
| | | Alive | 1 | 6 | 2 | 15 | 0 | 2 | 7 | 7 |

### *1.3.2 Outcome After Head Injury*

Table 1.1 reports data presented by Titterington et al. (1981) in a study initiated by the Institute of Neurological Sciences in Glasgow. These data were collected prospectively by neurosurgeons between 1968 and 1976. The original aim was to predict recovery for individual patients on the basis of data collected shortly after the injury. The data that we consider contain information on a binary outcome, $Y = 0/1$, corresponding to dead/alive after head injury, and the covariates: pupils (with good corresponding to a reaction to light and poor to no reaction), coma score (representing depth of coma, low or high), hematoma present (no/yes), and age (categorized as 1–25, 26–54, $\geq$55).

The response of interest here is $p(\boldsymbol{x}) = \Pr(Y = 1 \mid \boldsymbol{x})$; the probability that a patient with covariates $\boldsymbol{x}$ is alive. This quantity must lie in the range [0,1], and so, at least in this respect, linear models are unappealing. To illustrate, suppose we have a univariate continuous covariate $x$ and the model

$$p(x) = \beta_0 + \beta_1 x.$$

While probabilities not close to zero or one may change at least approximately linearly with $x$, it is extremely unlikely that this behavior will extend to the extremes, where the probability–covariate relationship must flatten out in order to remain in the correct range. An additional, important, consideration is that linear models commonly assume that the variance is constant and, in particular, does not depend on the mean. For a binary outcome with probability of response $p(x)$, the Bernoulli variance is $p(x)[1 - p(x)]$ and so depends on the mean. As we will see, accurate inference depends crucially on having modeled the mean–variance relationship appropriately.

A common model for binary data is the logistic regression model, in which the odds of death, $p(x)/[1 - p(x)]$, is modeled as a function of $x$. For example, the linear logistic regression model is

$$\frac{p(x)}{1 - p(x)} = \exp(\beta_0 + \beta_1 x).$$

This form is mathematically appealing, since the modeled probabilities are constrained to lie within [0,1], though the interpretation of the parameters $\beta_0$ and $\beta_1$ is not straightforward.

*Distinguishing Features.* Chapter 7 is dedicated to the modeling of binary data. In this chapter, logistic regression models are covered in detail, along with alternatives. Formulating predictive models and assessing the predictive power of such models is considered in Chaps. 10–12.

### *1.3.3 Lung Cancer and Radon*

We now describe an example in which the data arise from a spatial ecological study. In an ecological study, the unit of analysis is the group rather than the individual. In spatial epidemiological studies, due primarily to reasons of confidentiality, data on disease, population, and exposure are often available as aggregates across area. It is these areas that constitute the (ecological) group level at which the data are analyzed. In this example, we examine the association between lung cancer incidence (over the years 1998–2002) and residential radon at the level of the county, in Minnesota. Radon is a naturally occurring radioactive gas produced by the breakdown of uranium in soil, rock, and water and is a known carcinogen for lung cancer (Darby et al. 2001). However, in many ecological studies, when the association between lung cancer incidence and residential radon is estimated, radon appears protective. *Ecological bias* is an umbrella term that refers to the distortion of individual-level associations due to the process of aggregation. There are many facets to ecological bias (Wakefield 2008), but an important issue in the lung cancer/radon context is the lack of control for confounding, a primary source being smoking.

Let $Y_i$ denote the lung cancer incidence count and $x_i$ the average radon in county $i = 1, \ldots, n = 87$. Age and gender are strongly associated with lung cancer incidence, and a standard approach to controlling these factors is to form *expected counts* $E_i = \sum_{j=1}^{J} N_{ij} q_j$ in which we multiply the population in stratum $j$ and county $i$, $N_{ij}$, by a "reference" probability of lung cancer in stratum $j$, $q_j$, to obtain the expected count in stratum $j$. Summing over all $J$ stratum gives the total expected count. Intuitively, these counts are what we would expect if the disease rates in county $i$ conform with the reference. A summary response measure in county $i$ is the standardized morbidity ratio (SMR), given by $Y_i/E_i$. Counties with SMRs greater than 1 have an excess of cases, when compared to that expected.

Figure 1.3 maps the SMRs in counties of Minnesota, and we observe more than twofold variability with areas of high incidence in the northeast of the state. Figure 1.4 maps the average radon by county, with low radon in the counties to the northeast. This negative association is confirmed in Fig. 1.5 in which we plot the SMRs versus average radon, with a smoother indicating the local trend.

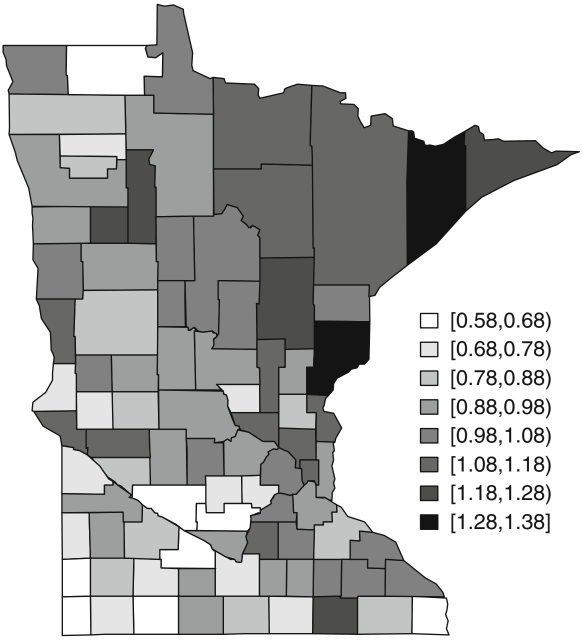

**Fig. 1.3** Standardized morbidity ratios for lung cancer in the period 1998–2002 by county in Minnesota

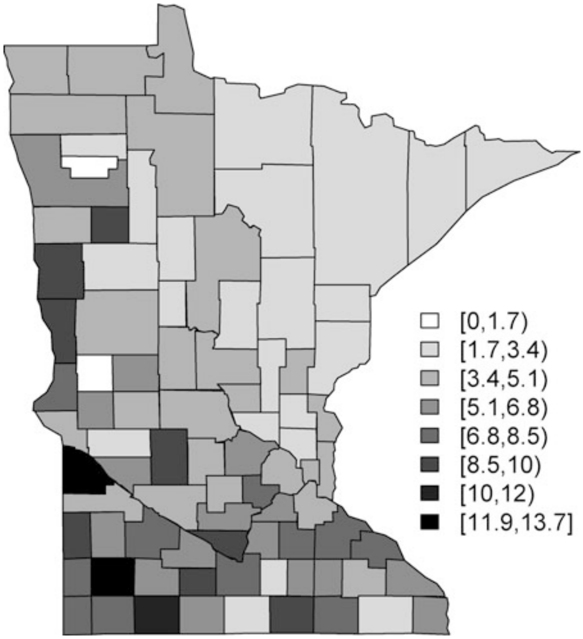

**Fig. 1.4** Average radon (pCi/liter) by county in Minnesota

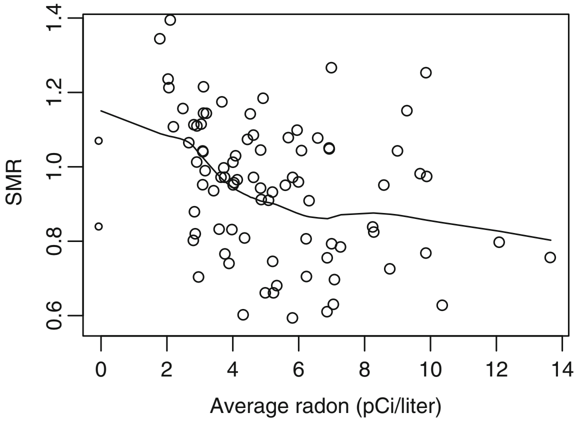

**Fig. 1.5** Standardized morbidity ratios versus average radon (pCi/liter) by county in Minnesota

A simple model that constrains the mean to be positive is the *loglinear regression*

$$\log \mathrm{E}\left[ \frac{Y_i}{E_i} \middle| x_i \right] = \beta_0 + \beta_1 x_i$$

$i = 1, \dots, n$. We might combine this form with a Poisson model for the counts. However, in a Poisson model, the variance is constrained to equal the mean, which is often too restrictive in practice, since excess-Poisson variability is often encountered. Hence, we would prefer to fit a more flexible model. We might also be concerned with residual spatial dependence between disease counts in counties that are close to each other. Information on confounder variables, especially smoking, would also be desirable.

*Distinguishing Features.* Poisson regression models for independent data, and extensions to allow for excess-Poisson variation, are described in Chap. 6. Such models are explicitly designed for nonnegative response variables. Accounting for residual spatial dependence is considered in Chap. 9.

### *1.3.4 Pharmacokinetic Data*

Pharmacokinetics is the study of the time course of a drug and its metabolites after introduction into the body. A typical experiment consists of a known dose of drug being administered via a particular route (e.g., orally or via an injection) at a known time. Subsequently, blood samples are taken, and the concentration of the drug is measured. The data are in the form of $n$ pairs of points $[x_i, y_i]$, where $x_i$ denotes the sampling time at which the $i$th blood sample is taken and $y_i$ denotes the $i$th measured concentration, $i = 1, \dots, n$. We describe in some detail some of the contextual scientific background in order to motivate a particular regression model.

A typical dataset, taken from Upton et al. (1982), is tabulated in Table 1.2 and plotted in Fig. 1.6. These data were collected after a subject was given an oral dose

**Table 1.2** Concentration ($y$) of the drug theophylline as a function of time ($x$), obtained from a subject who was administered an oral dose of size 4.53 mg/kg

| Observation<br>number<br>$i$ | Time<br>(hours)<br>$x_i$ | Concentration<br>(mg/liter)<br>$y_i$ |
|---|---|---|
| 1 | 0.27 | 4.40 |
| 2 | 0.58 | 6.90 |
| 3 | 1.02 | 8.20 |
| 4 | 2.02 | 7.80 |
| 5 | 3.62 | 7.50 |
| 6 | 5.08 | 6.20 |
| 7 | 7.07 | 5.30 |
| 8 | 9.00 | 4.90 |
| 9 | 12.15 | 3.70 |
| 10 | 24.17 | 1.05 |

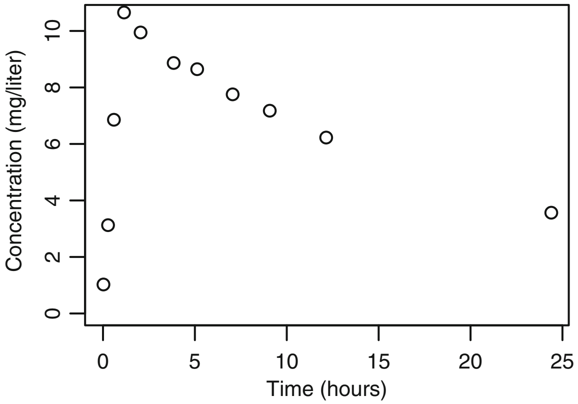

**Fig. 1.6** Concentration of theophylline plotted versus time for the data of Table 1.2

of 4.53 mg/kg of the antiasthmatic agent theophylline. The concentration of drug was determined in subsequent blood samples using a chemical assay (a method for determining the amount of a specific substance in a sample). Data were collected over a period slightly greater than 24 h following drug administration.

Pharmacokinetic experiments are important as they help in understanding the absorption, distribution, and elimination processes of drugs. Such an understanding provides information that may be used to decide upon the sizes and timings of doses that should be administered in order to achieve concentrations falling within a desired therapeutic window. Often the concentration of drug acts as a surrogate for the therapeutic response. The aim of a pharmacokinetic trial may be dose recommendation for a specific population, for example, to determine a dose size for the packaging, or recommendations for a particular patient based on covariates, which is known as *individualization*. A typical question is, for the patient who produced the data in Table 1.2, what dose could we give at 25 h to achieve a concentration of 10 mg/l at 37 h?

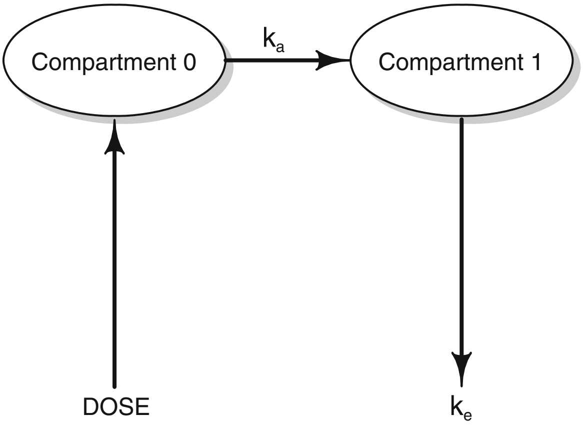

**Fig. 1.7** Representation of a one-compartment system with oral dosing. Concentrations are measured in compartment 1

The processes determining drug concentrations are very complicated, but simple compartmental models (e.g., Godfrey 1983) have been found to mimic the concentrations observed in patients. The basic idea is to model the body as a system of compartments within each of which the kinetics of the drug flow is assumed to be similar. We consider the simplest possible model for modeling drug concentrations following the administration of an oral dose. The model is represented in Fig. 1.7 and assumes that the body consists of a compartment into which the drug is introduced and from which absorption occurs into a second "blood compartment." The compartments are labeled retrospectively as 0 and 1 in Fig. 1.7. Subsequently, elimination from compartment 1 occurs with blood samples taken from this compartment.

We now describe in some detail the one-compartment model with first-order absorption and elimination. Let $w_k(t)$ represent the amount of drug in compartment $k$ at time $t$, $k = 0, 1$. The drug flow between the compartments is described by the differential equations

$$\frac{dw_0}{dt} = -k_a w_0, \tag{1.2}$$

$$\frac{dw_1}{dt} = k_a w_0 - k_e w_1, \tag{1.3}$$

where $k_a > 0$ is the absorption rate constant associated with the flow from compartment 0 to compartment 1 and $k_e > 0$ is the elimination rate constant (see Fig. 1.7). At time zero, the initial dose is $w_0(0) = D$, and solving the pair of differential equations (1.2) and (1.3), subject to this condition, gives the amount of drug in the body at time $x$ as

$$w_1(x) = \frac{Dk_a}{k_a - k_e}\left[\exp(-k_e x) - \exp(-k_a x)\right]. \tag{1.4}$$

We do not measure the amount of total drug but drug concentration, and so we need to normalize (1.4) by dividing $w_1(x)$ by the volume $V > 0$ of the blood compartment to give

$$\mu(x) = \frac{Dk_a}{V(k_a - k_e)} [\exp(-k_e x) - \exp(-k_a x)]. \tag{1.5}$$

so that $\mu(x)$ is the drug concentration in the blood compartment at time $x$. Equation (1.5) describes a model that is nonlinear in the parameters $V$, $k_a$ and $k_e$; for reasons that will be examined in detail in Chap. 6, inference for such models is more difficult than for their linear counterparts.

We have so far ignored the stochastic element of the model. An obvious error model is

$$y_i = \mu(x_i) + \epsilon_i,$$

with $\text{E}[\epsilon_i] = 0$, $\text{var}(\epsilon_i) = \sigma_\epsilon^2$, $i = 1, \ldots, n$, and $\text{cov}(\epsilon_i, \epsilon_j) = 0$, $i \neq j$. We may go one stage further and assume $\epsilon_i \mid \sigma_\epsilon^2 \sim_{iid} \text{N}(0, \sigma_\epsilon^2)$ where $\sim_{iid}$ is shorthand for "is independent and identically distributed as." There are a number of potential difficulties with this error model, beyond the distributional choice of normality. Concentrations must be nonnegative, and so we might expect the magnitude of errors to decrease with decreasing "true" concentration $\mu(x)$, a phenomenon that is often confirmed by examination of assay validation data. The error terms are likely to reflect not only assay precision, however, but also model misspecification, and given the simple one-compartment system we have assumed, this could be substantial. We might therefore expect the error terms to display correlation across time. In this example, the scientific context therefore provides not only a mean function but also information on how the variance of the data changes with the mean.

One simple solution, to at least some of these difficulties, is to take the logarithm of (1.5) and fit the model:

$$\log y_i = \log \mu(x_i) + \delta_i.$$

We may further assume $\text{E}[\delta_i] = 0$, $\text{var}(\delta_i) = \sigma_\delta^2$, $i = 1, ..., n$, and $\text{cov}(\delta_i, \delta_j) = 0$, $i \neq j$, multiplicative errors on the original scale and additive errors on the log scale give

$$\text{var}(Y) = \mu(x)^2 \text{var}(\text{e}^\delta) \approx \mu(x)^2 \sigma_\delta^2$$

for small $\delta$.

There are two other issues that are relevant to modeling in this example. The first is that in pharmacokinetic analyses, interest often focuses on *derived* parameters of interest, which are functions of $[V, k_a, k_e]$. In particular, we may wish to make inference for the time to maximum concentration, the maximum concentration, the clearance (initial dose divided by the area under the concentration curve), and the elimination half-life, which are given by

$$x_{\max} = \frac{1}{k_a - k_e} \log\left(\frac{k_a}{k_e}\right)$$

$$\begin{aligned} c_{\max} &= \mu(x_{\max}) = \frac{D}{V}\left(\frac{k_e}{k_a}\right)^{k_e/(k_a-k_e)} \\ \text{Cl} &= V \times k_e \\ t_{1/2} &= \frac{\log 2}{k_e}. \end{aligned}$$

A second issue is that model (1.5) is unidentifiable in the sense that the parameters $[V, k_a, k_e]$ give the same curve as the parameters $[Vk_e/k_a, k_e, k_a]$. This identifiability problem can be overcome via a restriction such as constraining the absorption rate to exceed the elimination rate, $k_a > k_e > 0$, though this complicates inference.

Often the data available for individualization will be sparse. For example, suppose we only observed the first two observations in Table 1.2. In this situation, inference is impossible without additional information (since there are more parameters than data points), which suggests a Bayesian approach in which prior information on the unknown parameters is incorporated into the analysis.

*Distinguishing Features.* Model (1.5) is nonlinear in the parameters. Such models will be considered in Chap. 6, including their use in situations in which additional information on the parameters is incorporated via the specification of a prior distribution. The data in Table 1.2 are from a single subject. In the original study, data were available for 12 subjects, and ideally we would like to analyze the totality of data; hierarchical models provide one framework for such an analysis. Hierarchical nonlinear models are considered in Chap. 9.

### 1.3.5 Dental Growth

Table 1.3 gives dental measurements of the distance in millimeters from the center of the pituitary gland to the pteryo-maxillary fissure in 11 girls and 16 boys recorded at the ages of 8, 10, 12, and 14 years. These data were originally analyzed in Potthoff and Roy (1964).

Figure 1.8 plots these data, and we see that dental growth for each child increases in an approximately linear fashion. Three inferential situations are:

1. *Summarization*. For each of the boy and girl populations, estimate the mean and standard deviation of pituitary gland measurements at each of the four ages.
2. *Population inference*. For each of the populations of boys and girls from which these data were sampled, estimate the average linear growth over the age range 8–14 years. Additionally, estimate the average dental distance, with an associated interval estimate, at an age of 9 years.
3. *Individual inference*. For a specific boy or girl in the study, estimate the rate of growth over the age range 8–14 years and predict the growth at 15 years. Additionally, for an unobserved girl, from the same population that produced the sampled girls, obtain a predictive growth curve, along with an interval envelope.

**Table 1.3** Dental growth data for boys and girls

| | Age (years) | | | | | Age (years) | | | |
|---|---|---|---|---|---|---|---|---|---|
| Girl | 8 | 10 | 12 | 14 | Boy | 8 | 10 | 12 | 14 |
| 1 | 21.0 | 20.0 | 21.5 | 23.0 | 1 | 26.0 | 25.0 | 29.0 | 31.0 |
| 2 | 21.0 | 21.5 | 24.0 | 25.5 | 2 | 21.5 | 22.5 | 23.0 | 26.5 |
| 3 | 20.5 | 24.0 | 24.5 | 26.0 | 3 | 23.0 | 22.5 | 24.0 | 27.5 |
| 4 | 23.5 | 24.5 | 25.0 | 26.5 | 4 | 25.5 | 27.5 | 26.5 | 27.0 |
| 5 | 21.5 | 23.0 | 22.5 | 23.5 | 5 | 20.0 | 23.5 | 22.5 | 26.0 |
| 6 | 20.0 | 21.0 | 21.0 | 22.5 | 6 | 24.5 | 25.5 | 27.0 | 28.5 |
| 7 | 21.5 | 22.5 | 23.0 | 25.0 | 7 | 22.0 | 22.0 | 24.5 | 26.5 |
| 8 | 23.0 | 23.0 | 23.5 | 24.0 | 8 | 24.0 | 21.5 | 24.5 | 25.5 |
| 9 | 20.0 | 21.0 | 22.0 | 21.5 | 9 | 23.0 | 20.5 | 31.0 | 26.0 |
| 10 | 16.5 | 19.0 | 19.0 | 19.5 | 10 | 27.5 | 28.0 | 31.0 | 31.5 |
| 11 | 24.5 | 25.0 | 28.0 | 28.0 | 11 | 23.0 | 23.0 | 23.5 | 25.0 |
| | | | | | 12 | 21.5 | 23.5 | 24.0 | 28.0 |
| | | | | | 13 | 17.0 | 24.5 | 26.0 | 29.5 |
| | | | | | 14 | 22.5 | 25.5 | 25.5 | 26.0 |
| | | | | | 15 | 23.0 | 24.5 | 26.0 | 30.0 |
| | | | | | 16 | 22.0 | 21.5 | 23.5 | 25.0 |

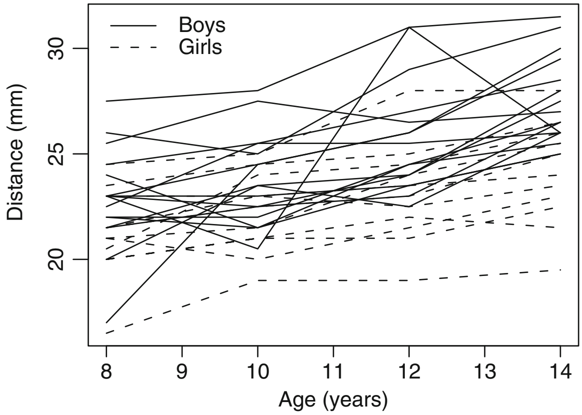

**Fig. 1.8** Dental growth data for boys and girls: distance plotted versus age

With 16 boys and 11 girls, inference for situation 1 can be achieved by simply evaluating the sample mean and standard deviation at each time point; these quantities are given in Table 1.4. These simple summaries are straightforward to construct and are based on independence of individuals. To obtain interval estimates for the means and standard deviations, one must be prepared to make assumptions (such as approximate normality of the measurements), since for these data the sample sizes are not large and we might be wary of appealing to large sample (asymptotic) arguments.

**Table 1.4** Sample means and standard deviations (SDs) for girls and boys, by age group

| Age | Girls | | Boys | |
|---|---|---|---|---|
| (years) | Mean (mm) | SD (mm) | Mean (mm) | SD (mm) |
| 8 | 21.2 | 2.1 | 22.9 | 2.5 |
| 10 | 22.2 | 1.9 | 23.8 | 2.1 |
| 12 | 23.1 | 2.4 | 25.7 | 2.7 |
| 14 | 24.1 | 2.4 | 27.5 | 2.1 |

For situation 2, we may fit a linear model relating distance to age. Since there are no data at 9 years, to obtain an estimate of the dental distance, we again require a model relating distance to age. In situation 3, we may wish to use the totality of data as an aid to providing inference for a specific child. For a new girl from the same population, we clearly need to use the existing data and a model describing between-girl differences.

For longitudinal (repeated measures) data such as these, we cannot simply fit models to the totality of the data on boys or girls and assume independence of measurements; we need to adjust for the correlation between measurements on the same child. There is clearly dependence between such measurements. For example, boy 10 has consistently higher measurements than the majority of boys. There are two distinct approaches to modeling longitudinal data. In the *marginal* approach, the average response is modeled as a function of covariates (including time), and standard errors are empirically adjusted for dependence. In the *conditional* approach, the response of each individual is modeled as a function of individual-specific parameters that are assumed to arise from a distribution, so that the overall variability is partitioned into within- and between-child components. The marginal approach is designed for estimating population-level questions (as posed in situation 2) based on minimal assumptions. Conditional approaches can answer a greater number of inferential questions but require an increased number of assumptions which decreases their robustness to model misspecification.

*Distinguishing Features.* Chapter 8 describes linear models for dependent data such as these.

## *1.3.6 Spinal Bone Mineral Density*

Bachrach et al. (1999) analyze longitudinal data on spinal bone mineral density (SBMD) measurements on 230 women aged between 8 and 27 years and of one of four ethnic groups: Asian, Black, Hispanic, and White. The aim of this study was to examine ethnic differences in SBMD.

Figure 1.9 displays the SBMD measurements by individual, with one panel for each of the four races. The relationship between SBMD and age is clearly nonlinear, and there are also woman-specific differences in overall level so that observations

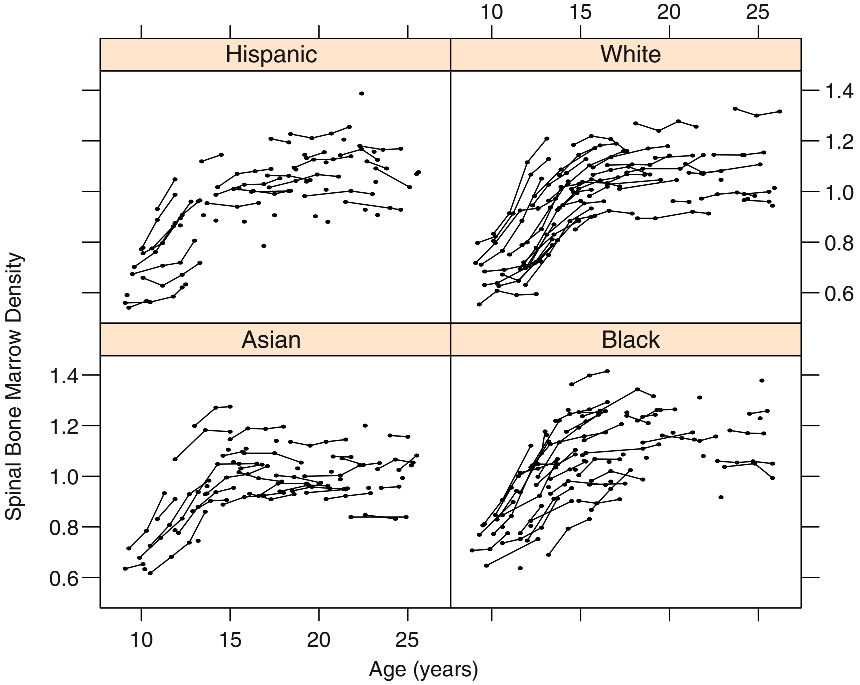

**Fig. 1.9** Spinal bone mineral density measurements as a function of age and ethnicity. Points that are connected represent measurements from the same woman

on the same woman are correlated. Letting $Y_{ij}$ represent the SBMD measurement on woman $i$ at age $\text{age}_{ij}$, we might propose a mean model of the form

$$\text{E}[Y_{ij} \mid \text{age}_{ij}] = \boldsymbol{x}_i \boldsymbol{\beta} + f(\text{age}_{ij}) + b_i$$

where $\boldsymbol{x}_i$ is a $1 \times 4$ row vector with a single one and three zeroes that represents the ethnicity of woman $i$ (coded in the order Hispanic, White, Asian, Black), with $\boldsymbol{\beta} = [\beta_H, \beta_W, \beta_A, \beta_B]^{\text{T}}$ the $4 \times 1$ vector of associated regression coefficients, $f(\text{age}_{ij})$ is a function that varies smoothly with age, and $b_i$ is a woman-specific intercept which is included to account for dependencies of measurements on the same individual. The relationship between SBMD and age is not linear and not of primary interest. Consequently, we would like to use a flexible model form, and we may not be concerned if this model does not contain easily interpretable parameters. Nonparametric regression is the term we use to refer to flexible mean modeling.

*Distinguishing Features.* The analysis of these data requires both a flexible mean model for the age effect and acknowledgement of the dependence of measurements on the same woman. Chapters 10–12 describe models that allow for these possibilities.

## 1.4 Nature of Randomness

Regression models consist of both deterministic and stochastic (random) components, and a consideration of the sources of the randomness is worthwhile, both to interpret parameters contained in the deterministic component and to model the stochastic component. We initially consider an idealized situation in which a response is completely deterministic, given sufficient information, and randomness is only induced by missing information.[3] Let $y$ denote a variable with values $y_1, \dots, y_N$ within a population. We begin with a very simple deterministic model

$$y_i = \beta_0 + \beta_1 x_i + \gamma z_i \tag{1.6}$$

for $i = 1, \dots, N$, so that, given $x_i$ and $z_i$ (and knowing $\beta_0, \beta_1$ and $\gamma$), $y_i$ is completely determined. Suppose we only measure $y_i$ and $x_i$ and assume the model

$$Y_i = \beta_0^\star + \beta_1^\star x_i + \epsilon_i.$$

To interpret $\beta_0^\star$ and $\beta_1^\star$, we need to understand the relationship between $x_i$ and $z_i$, $i = 1, \dots, N$. To this end, write

$$z_i = a + b x_i + \delta_i, \tag{1.7}$$

$i = 1, \dots, N$. This form does not in any sense assume that a linear association is appropriate or "correct", rather it is the linear approximation to $\text{E}[Z \mid x]$. In (1.7), we may take $a$ and $b$ as the least squares estimates from fitting a linear model to the data $[x_i, z_i]$, $i = 1, \dots, N$. Substitution of (1.7) into (1.6) yields

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \gamma(a + b x_i + \delta_i) \\ &= \beta_0^\star + \beta_1^\star x_i + \epsilon_i \end{aligned}$$

where

$$\begin{aligned} \beta_0^\star &= \beta_0 + a\gamma \\ \beta_1^\star &= \beta_1 + b\gamma \\ \epsilon_i &= \gamma\delta_i, \qquad i = 1, \dots, N, \end{aligned} \tag{1.8}$$

[3] When simulations are performed, pseudorandom numbers are generated via deterministic sequences. For example, consider the sequence generated by the *congruential generator*

$$X_i = a X_{i-1}, \ \ \text{mod}(m)$$

along with initial value (or "seed") $X_0$. Then $X_i$ takes values in $0, 1, \dots, m-1$, and pseudorandom numbers are obtained as $U_i = X_i/m$, where $X_0$, $a$, and $m$ are chosen so that the $U_i$'s have (approximately) the properties of uniform $\text{U}(0, 1)$ random variables. However, if $X_0$, $a$, and $m$ are known, the randomness disappears! Ripley (1987, Chap. 2) provides a discussion of pseudorandom variable generation and specifically "good" choices of $a$ and $m$.

so that $\beta_1^\star$ is a combination of the direct effect of $x_i$ on $y_i$, *and* the effect of $z_i$, through the linear association between $z_i$ and $x_i$. This development illustrates the problems in nonrandomized situations of estimating the causal effect of $x_i$ on $y_i$, that is, $\beta_1$. Turning to the stochastic component (1.8) illustrates that properties of $\epsilon_i$ are inherited from $\delta_i$. Hence, assumptions such as constancy of variance of $\epsilon_i$ depend on the nature of $z_i$ and, in particular, on the joint distribution of $x_i$ and $z_i$.

Increasing slightly the realism, we extend the original deterministic model to

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{k=1}^{q} \gamma_k z_{ik}. \tag{1.9}$$

Suppose we only measure $x_{i1}, \dots, x_{ip}$ and assume the simple model

$$Y_i = \beta_0^\star + \sum_{j=1}^{p} \beta_j^\star x_{ij} + \epsilon_i, \tag{1.10}$$

where the errors, $\epsilon_i$, now correspond to the totality of scaled versions of the $z_{ik}$'s that remain after extracting the linear associations with the $x_{ij}$'s by analogy with (1.7) and (1.8).

Viewing the error terms as sums of random variables and considering the central limit theorem (Appendix G) naturally leads to the normal distribution as a plausible error distribution. There is no compelling reason to believe that the variance of this normal distribution will be constant across the space of the $x$ variables, however.

We have distinguished between the regression coefficients in the assumed model (1.10), denoted by $\beta_j^\star$, and those in the original model (1.9), denoted $\beta_j$. In general, $\beta_j \neq \beta_j^\star$, because of the possible effects of *confounding* which occurs due to dependencies between $x_{ij}$ and elements of $\boldsymbol{z}_i = [z_{i1}, \dots, z_{iq}]$. In the example just considered, only if $x_{ij}$ is linearly independent of the $z_{ik}$ will the coefficients $\beta_j$ and $\beta_j^\star$ coincide. For nonlinear models, the relationship between the two sets of coefficients is even more complex.

This development illustrates that an aim of regression modeling is often to "explain" the error terms using observed covariates. In general, error terms represent not only unmeasured variables but also data anomalies, such as inaccurate recording of responses and covariates, and model misspecification. Clearly the nature of the randomness, and the probabilities we attach to different events, is conditional upon the information that we have available and, specifically, the variables we measure.

Similar considerations can be given to other types of random variables. For example, suppose we wish to model a binary random variable $Y$ taking values coded as 0 and 1. Sometimes it will be possible to link $Y$ to an underlying continuous *latent* variable and use similar arguments to that above. To illustrate, $Y$ could be an indicator of low birth weight and is a simple function of the true birth weight, $U$, which is itself associated with many covariates. We may then model the probability of low birth weight as a function of covariates $\boldsymbol{x}$, via

$$p(\boldsymbol{x}) = \Pr(Y = 1 \mid \boldsymbol{x}) = \Pr(U \le u_0 \mid \boldsymbol{x}) = \mathrm{E}[Y \mid \boldsymbol{x}],$$

where $u_0$ is the threshold value that determines whether a child is classified as low birth weight or not. This development is taken further in Sects. 7.6.1 and 9.13.

The above gives one a way of thinking about where the random terms in models arise from, namely as unmeasured covariates. In terms of distributional assumptions, some distributions arise naturally as a consequence of simple physical models. For example, suppose we are interested in modeling the number of events occurring over time. The process we now describe has been found empirically to model a number of phenomena, for example the arrival of calls at a telephone exchange or the emission of particles from a radioactive source. Let the rate of occurrences be denoted by $\rho > 0$ and $N(t, t + \Delta t)$ be the number of events in the interval $(t, t + \Delta t]$. Suppose that, informally speaking, $\Delta t$ tends to zero from above and that

$$\begin{aligned} \Pr\left[N(t, t + \Delta t) = 0\right] &= 1 - \rho\Delta t + o(\Delta t), \\ \Pr\left[N(t, t + \Delta t) = 1\right] &= \rho\Delta t + o(\Delta t), \end{aligned}$$

so that $\Pr\left[N(t, t + \Delta t) > 1\right] = o(\Delta t)$. The notation $o(\Delta t)$ represents a function that tends to zero more rapidly than $\Delta t$. Finally, suppose that $N(t, t + \Delta t)$ is independent of occurrences in $(0, t]$. Then we have a *Poisson process*, and the number of events occurring in the fixed interval $(t, t + h]$ is a Poisson random variable with mean $\rho h$.

Other distributions are “artificial.” For example, a number of distributions arise as functions of normal random variables (such as Student’s t, Snedecor’s F, and chi-squared random variables) or may be dreamt up for flexible and convenient modeling (as is the case for the so-called Pearson family of distributions).

Models can arise from idealized views of the phenomenon under study, but then we might ask: “If we could measure absolutely everything we wanted to, would there be any randomness left?” In all but the simplest experiments, this question is probably not that practically interesting, but the central idea of quantum mechanics tells us that probability is still needed, because some experimental outcomes are fundamentally unpredictable (e.g., Feynman 1951).

## 1.5 Bayesian and Frequentist Inference

What distinguishes the field of statistics from the use of statistical techniques in a particular discipline is a principled approach to inference in the face of uncertainty. There are two dominant approaches to inference, which we label as Bayesian and frequentist, and each produces inferential procedures that are optimal with respect to different criteria.

In Chaps. 2 and 3, we describe, respectively, the frequentist and Bayesian approaches to statistical inference. Central to the philosophy of each approach is the interpretation of probability that is taken. In the frequentist approach, as the name suggests, probabilities are viewed as limiting frequencies under infinite hypothetical replications of the situation under consideration. Inferential recipes, such as specific estimators, are assessed with respect to their performance under repeated sampling of the data, with model parameters viewed as fixed, albeit unknown, constants. By contrast, in the Bayesian approach that is described in this book, probabilities are viewed as subjective and are interpreted conditional on the available information. As a consequence, assigned probabilities concerning the same event may differ between individuals. In this sense probabilities do not exist as they vary as a function of the available information. All unknown parameters in a model are treated as random variables, and inference is based upon the (posterior) probability distribution of these parameters, given the data and other available information. Practically speaking, the interpretation of probability is less relevant than the number of assumptions that are required for valid inference (which has implications for the robustness of analysis) and the breadth of inferential questions that can be answered using a particular approach.

It should be stressed that many issues arising in the analysis of regression data (such as the nature of the sampling scheme, parameter interpretation, and misspecification of the mean model) are independent of philosophy and in practice are usually of far greater importance than the inferential approach taken to analysis.

Each of the frequentist and Bayesian approaches have their merits and can often be used in tandem, an approach we follow and advocate throughout this book. If substantive conclusions differ between different approaches, then discovering the reasons for the discrepancies can be informative as it may reveal that a particular analysis is leaning on inappropriate assumptions or that relevant information is being ignored by one of the approaches. Those situations in which one of the approaches is more or less suitable will also be distinguished throughout this book, with a short summary being given in the next section.

## 1.6 The Executive Summary

I would like to briefly summarize my view on when to take Bayesian or frequentist approaches to estimation. As the examples throughout this book show, on many occasions, if one is careful in execution, both approaches to analysis will yield essentially equivalent inference. For small samples, the Bayesian approach with thoughtfully specified priors is often the only way to go because of the difficulty in obtaining well-calibrated frequentist intervals. An example of such a sparse data occasion is given at the end of Sect. 6.16. For medium to large samples, unless there is strong prior information that one wishes to incorporate, a robust frequentist approach using sandwich estimation (or quasi-likelihood if one has faith in the variance model) is very appealing since consistency is guaranteed under relatively mild conditions. For highly complex models (e.g., with many random effects), a Bayesian approach is often the most convenient way to formulate the model, and computation under the Bayesian approach is the most straightforward. The modeling of spatial dependence in Sect. 9.7 provides one such example in which the Bayesian approach is the simplest to implement. The caveat to complex modeling is that in most cases consistency of inference is only available if all stages of the model are correctly specified. Consequently, if one really cares about interval estimates, then extensive model checking will be necessary. If formal inference is not required but rather one is in an exploratory phase, then there is far greater freedom to experiment with the approaches that one is most familiar with, including nonparametric regression. In this setting, using procedures that are less well-developed statistically is less dangerous.

In contrast to estimation, hypothesis testing using frequentist and Bayesian methods can often produce starkly differing results, even in large samples. As discussed in Chap. 4, I think that hypothesis testing is a very difficult endeavor, and tests applied using the frequentist approach, as currently practiced (with $\alpha$ levels being fixed regardless of sample size), can be very difficult to interpret. In general, I prefer estimation to hypothesis testing.

As a final comment, as noted, in many instances carefully conducted frequentist and Bayesian approaches will lead to similar substantive conclusions; hence, the choice between these approaches can often be based on that which is most natural (i.e., based on training and experience) to the analyst. Consequently, throughout this book, methods are discussed in terms of their advantages and shortcomings, but a strong recommendation of one method over another is usually not given as there is often no reason for stating a preference.

## 1.7 Bibliographic Notes

Rosenbaum (2002) provides an in-depth discussion of the analysis of data from observational studies, and an in-depth treatment of causality is the subject of Pearl (2009). A classic text on survey sampling is Cochran (1977) with Korn and Graubard (1999) and Lumley (2010) providing more recent presentations. Regression from a survey sampling viewpoint is discussed in the edited volume of Chambers and Skinner (2003). Errors-in-variables is discussed in detail by Carroll et al. (2006) and missing data by Little and Rubin (2002). Johnson et al. (1994, 1995, 1997); Kotz et al. (2000), and Johnson et al. (2005) provide a thorough discussion of the genesis of univariate and multivariate discrete and continuous probability distributions and, in particular, their relationships to naturally occurring phenomena. Barnett (2009) provides a discussion of the mechanics and relative merits of Bayesian and frequentist approaches to inference; see also Cox (2006).